

Multithreshold Entropy Linear Classifier

Wojciech Marian Czarnecki, Jacek Tabor

Faculty of Mathematics and Computer Science,

Jagiellonian University, Krakow, Poland.

{wojciech.czarnecki, jacek.tabor}@uj.edu.pl

August 6, 2014

Abstract

Linear classifiers separate the data with a hyperplane. In this paper we focus on the novel method of construction of multithreshold linear classifier, which separates the data with multiple parallel hyperplanes. Proposed model is based on the information theory concepts – namely Renyi’s quadratic entropy and Cauchy-Schwarz divergence.

We begin with some general properties, including data scale invariance. Then we prove that our method is a multithreshold large margin classifier, which shows the analogy to the SVM, while in the same time works with much broader class of hypotheses. What is also interesting, proposed method is aimed at the maximization of the balanced quality measure (such as Matthew’s Correlation Coefficient) as opposed to very common maximization of the accuracy. This feature comes directly from the optimization problem statement and is further confirmed by the experiments on the UCI datasets.

It appears, that our Entropy Multithreshold Linear Classifier (MELC) obtains similar or higher scores than the ones given by SVM on both synthetic and real data. We show how proposed approach can be beneficial for the cheminformatics in the task of ligands activity prediction, where despite better classification results, MELC gives some additional insight into the data structure (classes of underrepresented chemical compounds).

1 Introduction

Linear classifiers (SVM, perceptron, LDA, logistic regression) aim to find $v \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that the decision on the class of x is based on

$$\text{sign}(v^T x - b). \quad (1)$$

The linear classification is important as it has the advantage of small VC dimension. The same ideas can be seen behind the neural networks (and their modifications like Extreme Learning Machines [1] or Deep Learning [2]), where

the activation of the single neuron is given by (1), while the role played by it in the whole decision process is usually given by

STEP 1: calculate $v^T x$,

STEP 2: make decision based on the sign of $v^T x - b$.

Although the linear classification is usually very efficient, even for the simple sets in \mathbb{R} , like $+-+$, see Figure 1, we cannot obtain sufficient classification results. This led to the need for kernelization procedure [3].

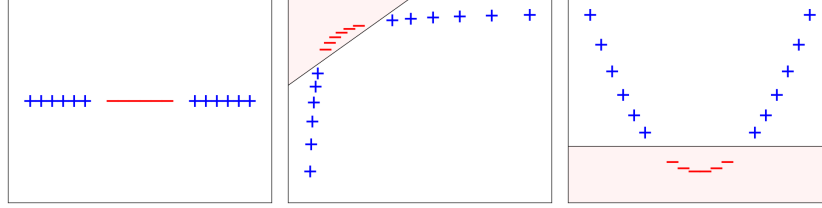


Figure 1: From left: $+-+$ dataset (linearly non-separable), $+-+$ dataset in trained neural network with 2 hidden nodes with sigmoid activation functions, $+-+$ dataset in trained SVM with polynomial kernel of degree 2

Our postulate is that by applying the second step we often lose some of the information given by the first one – observe that both in $+-+$ or XOR case we can make sufficiently good classification decision based on the knowledge of the value of $v^T x$ (for well chosen v), see Figure 6. One can therefore ask why we do not use the additional information? One of the possible answers lies in the fact that most classification methods, like SVM, aim at building a “large” linear margin between classes, which in a natural way leads to the single-threshold decision boundary.

Thus there appears a natural question if we can construct a classification method which would find the projection $x \rightarrow v^T x$ which could directly deal with more complex classification cases like $+-+$ and XOR. The problem in fact splits into two – how to find the right $v \in \mathbb{R}^d$ and how to make the proper classification decision in \mathbb{R} . The answer for the second question is given by multithreshold linear classifiers [4], where instead of decision based on the split of \mathbb{R} into $(-\infty, b)$ and $[b, \infty)$ the division into finite number of intervals is allowed¹.

The answer to the first question is nontrivial, and in our opinion there could be many reasonable solutions. In this paper we have decided to base the decision on entropy-based divergence measure [5]. We have chosen the *Renyi's quadratic entropy*

$$H_2(f) := -\log \int f^2$$

¹This type of classification can be obtain in particular by the density based classifiers in \mathbb{R} .

and the connected *Cauchy-Schwarz divergence*

$$\begin{aligned} D_{CS}(f, g) &:= \log \int f^2 + \log \int g^2 - 2 \log \int fg \\ &= -(H_2(f) + H_2(g) + 2 \log \dot{p}^\times(f, g)), \end{aligned} \quad (2)$$

where $\dot{p}^\times(f, g) := \int fg$ denotes the *cross-information potential*. Our reasons behind such a choice are the following:

- Renyi entropy and the Cauchy-Schwarz divergence are easily computable and the exact formulas for the Gaussian mixtures are known (this allows the use of gradient methods in our optimization problem, see Practical Considerations Section),
- the Cauchy-Schwarz divergence is translation and scale invariant in terms of input data transformation,
- D_{CS} has nice theoretical properties, as the minimization of \dot{p}^\times leads to the maximization of the multi-threshold boundary², while the part consisting of Renyi's entropies adds the regularizing term, see Theory Section.

From the practical point of view, we first project the data by v^T onto \mathbb{R} , and apply there the classical kernel density estimation given for the dataset $P \subset \mathbb{R}$ by

$$[P]_\sigma := \frac{1}{|P|} \sum_{p \in P} \mathcal{N}(p, \sigma^2), \quad (3)$$

where $\mathcal{N}(p, \sigma^2)$ denotes the one dimensional normal distribution. We skip the subscript σ (which denotes the window width) if it is chosen according to the Silverman's rule [6]

$$\sigma = (4/3)^{1/5} |P|^{-1/5} \sigma_P, \quad (4)$$

where σ_P denotes the standard deviation of the data P . Only later we calculate the Cauchy-Schwarz divergence. It is important to notice that our method performs density estimation in one-dimensional space \mathbb{R} . It is a common knowledge that density estimation in high dimensions is unreliable (requires enormous amount of samples), which is one of the reasons why purely density based classification is rarely used. In particular, even in the simplest case when data comes from multivariate normal distribution and we are interested in the good estimation of the value at 0, we need over 10,000 samples for just 7 dimensions [6]. On the other hand, in one dimension we just need 4 samples (to obtain a solution with 0.1 precision in terms of mean squared error). This supports the idea behind creation of models based on 1-dimensional linear projections as they provide reliable estimation of the underlying densities.

Consequently our final optimization problem can be formulated as follows:

Optimization problem. Consider classes X_+ and X_- in \mathbb{R}^d . Find nonzero $v \in \mathbb{R}^d$ which maximizes the value of

$$D_{CS}([v^T X_+], [v^T X_-]).$$

²To some extent we obtain multi-threshold analogue of large margin classifier.

The resulting multithreshold classifier is constructed from the density estimations $\llbracket v^T X_+ \rrbracket$ and $\llbracket v^T X_- \rrbracket$. Observe that, contrary to SVM, in our basic method we do not have any free parameters.

As it is shown in the Evaluation Section, such model usually obtains similar or better classification quality than the linear SVM. It occurs that in practice due to the strong regularization proposed method selects quite small number of thresholds (which reduces the VC dimension [7] of the resulting model). In fact, when using Silverman’s rule for kernel window width estimation, our method built a single threshold model in nine out of ten UCI datasets. It is worth noting that these solutions are significantly different from the ones given by SVM so, even though their scores are similar, proposed method is fundamentally different and therefore gives additional knowledge of the problem.

The interesting practical applications of multithreshold model is the more detailed insight into data geometry. Let us consider the task of ligands activity prediction for given proteins (which is further described in the Evaluation Section). Figure 2 shows results of kernel density estimation for one of the obtained models for cathepsin ligands [8]. One can notice how multithreshold classifier

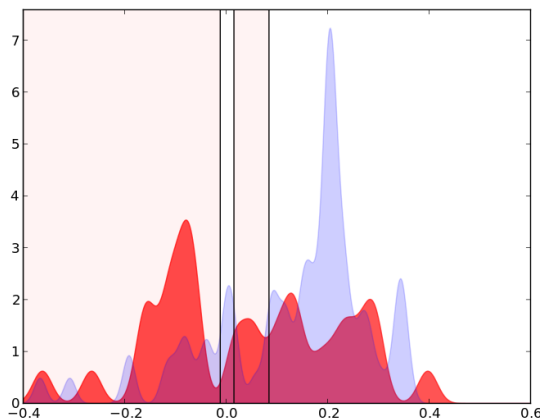


Figure 2: Kernel density estimation of the linear projection of one of the folds of cathepsin ligands detection task using proposed multithreshold linear classifier for the test set. The negative class spans through x such that $v^T x \in (\infty, -0.02] \cup (0.02, 0.09]$ and the positive one through x such that $v^T x \in (-0.02, 0.02] \cup (0.09, \infty)$

exploits the internal structure of the data by capturing small group of data points which is a part of the different class which would be ignored in linear classification. This results in the significant increase in the classification quality compared to the commonly encountered in this domain SVM model. This shows how the proposed model is able to exploit additional knowledge from the simple linear data projection. In the case of cheminformatics domain this typically rep-

resents some specific group of compounds³, distinctive from the most popular active ones (positive samples) and therefore is especially worth investigation⁴.

To sum up the MELC (Multithreshold Entropy Linear Classifier) has the following advantages:

- has strong theoretical background based on Information Theory,
- can build both single- and multithreshold linear classifiers,
- maximizes the balanced quality measure (is class imbalance invariant),
- is scale invariant (requires no data scaling),
- directly gives not only classification but also its likelihood (without the need for Platt’s scaling),
- behaves well as the parameter-free model,
- although it tries to maximize the margins it builds significantly different model than SVM,
- can be parametrized to better fit data, and this free parameter has clear geometrical intuition.

Its current biggest drawback is computational complexity and existence of local solutions.

Let us now briefly describe the contents of the paper. After short analysis of related work we show the basic properties of Cauchy-Schwarz divergence including its scale invariance and solutions for normally distributed data. Next, we prove that proposed model maximizes the margins’ sizes of multithreshold linear classifier and that the entropy terms play the regularization role. Then we proceed to some practical considerations regarding optimization procedure, its implementation and possible drawbacks. We conclude with the evaluation based on both synthetic and real datasets.

2 Related work

Multithreshold linear classifiers are present in machine learning for a long time [9, 10], however they did not receive as much attention as the single threshold ones. One of the reasons may be hardness of their theoretical analysis and lack of answers for very basic question like their exact Vapnik-Chervonenkis dimension [11]. On the other hand Anthony et al. [7] recently showed some bounds regarding this class of models. However, efficient training of such models remains an open issue [11].

³The main aim of this kind of research is identifying new drugs and 2 compounds which are biologically active.

⁴Exploiting such underrepresented groups of molecules might shed light on the currently under researched structural classes and lead to discovery of new types of drugs.

As we will show, our method is strongly related to the Support Vector Machines concept [3] or more generally large margin classifier idea [12, 13, 14]. It is worth noting that we are not presenting a modification of SVM model (dozens of which appeared in recent years) but rather propose a conceptually different approach which leads to some important similarities.

Renyi's entropy has been deeply analyzed in the recent book by Principe et al. [5], showing its wide applications spanning from classification optimization criterion [15], through clustering techniques [16] to ICA and other self-organizing methods [17]. Use of Cauchy-Schwarz divergence for the classification criterion has been investigated in the past, in particular for a simple multilayer neural networks [18]. However, to the authors best knowledge, it has not yet been used as a criterion for the choice of one-dimensional linear projection used for density-based classification.

In the broader sense, we are employing techniques from the information theory, which have been applied for construction of Decision Trees and, very successful model from 2001, Random Forest [19]. On the other hand, density estimation based models have been recently used as the base of Deep Learning architectures [2] and proved to be a very good data processing technique.

3 Cauchy-Schwarz divergence

In this section we discuss the basic theoretical aspects of the the Cauchy-Schwarz divergence. We show that the it is insensitive to the change of scale, which consequently yields that that we can restrict search to the unit sphere $S := \{v \in \mathbb{R}^d : \|v\| = 1\}$. Next we discuss the case of normal distributions.

3.1 Scale invariance

We are going to show that the Cauchy-Schwarz divergence is scale invariant. Observe that for $f, g : \mathbb{R} \rightarrow \mathbb{R}_+$

$$D_{CS}(f, g) = -2 \log \left(\int \frac{f}{\|f\|_2} \frac{g}{\|g\|_2} \right),$$

where $\|f\|_2$ denotes the L^2 -norm of f . This implies that

$$D_{CS}(\alpha f, \beta g) = D_{CS}(f, g) \text{ for } \alpha, \beta > 0,$$

which means that in the use of the Cauchy-Schwarz divergence we do not have to normalize the data.

We show that D_{CS} does not depend on the change of scale. To do so we need the following notation: for density f in \mathbb{R} , we put

$$R_\alpha f(x) := \frac{1}{|\alpha|} f(x/\alpha).$$

Observe that if $P \subset \mathbb{R}$ comes from the density f , then αP was generated from the density $R_\alpha f$. In other words the operation R_α corresponds (for the densities) to the operation of rescaling the data by $\alpha \neq 0$.

Lemma 3.1. *Consider densities f, g in \mathbb{R} and $\alpha \neq 0$. Then*

$$D_{CS}(f, g) = D_{CS}(R_\alpha f, R_\alpha g).$$

Proof. One can easily see that

$$\begin{aligned} \int R_\alpha h(x) R_\alpha \tilde{h}(x) dx &= \frac{1}{\alpha^2} \int h(x/\alpha) \tilde{h}(x/\alpha) dx \\ &\stackrel{u=x/\alpha}{=} \frac{1}{|\alpha|} \int h(u) \tilde{h}(u) du. \end{aligned}$$

Applying the above we obtain that

$$\begin{aligned} D_{CS}(R_\alpha f, R_\alpha g) &= \log \int (R_\alpha f)^2 + \log \int (R_\alpha g)^2 - 2 \log \int R_\alpha f R_\alpha g \\ &= \log \int f^2 - \log |\alpha| + \log \int g^2 - \log |\alpha| - 2 \log \int fg + 2 \log |\alpha| \\ &= D_{CS}(f, g). \end{aligned}$$

□

We obtain the following corollary as a direct consequence of the previous lemma and the fact that $[\![\alpha P]\!]_{\alpha r} = R_\alpha [\![P]\!]_r$.

Corollary 3.1. *Let $P_+, P_- \subset \mathbb{R}$ be given. Then*

$$D_{CS}([\![\alpha P_+]\!]_{\alpha r}, [\![\alpha P_-]\!]_{\alpha s}) = D_{CS}([\![P_+]\!]_r, [\![P_-]\!]_s).$$

Since $\sigma_{\alpha P} = |\alpha| \sigma_P$, we obtain by the Silverman's rule (4) that

$$[\![\alpha P]\!] = R_\alpha [\![P]\!] \text{ for } P \subset \mathbb{R}.$$

This implies that the Cauchy-Schwarz divergence of the data projection does not depend on the rescaling of the data:

$$D_{CS}([\![v^T(\alpha X_+)]\!], [\![v^T(\alpha X_-)]\!]) = D_{CS}([\![v^T X_+]\!]_r, [\![v^T X_-]\!]_s) \quad (5)$$

for $v \in \mathbb{R}^d$, $\alpha \neq 0$. Consequently, in its maximization process we can restrict to the unit sphere.

Finally, we arrive at:

Theorem 3.1. *Let $A : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a linear invertible map. Then*

$$\begin{aligned} &\sup\{D_{CS}([\![v^T(AX_+)]\!], [\![v^T(AX_-)]\!]) : v \in S\} \\ &= \sup\{D_{CS}([\![v^T X_+]\!]_r, [\![v^T X_-]\!]_s) : v \in S\}. \end{aligned}$$

Proof. Let $v \neq 0$ be arbitrarily fixed and let $w = A^T v$. Then

$$v^T(AX) = (A^T v)^T X = w^T X,$$

which implies that

$$D_{CS}(\llbracket v^T (AX_+) \rrbracket, \llbracket v^T (AX_-) \rrbracket) = D_{CS}(\llbracket w^T X_+ \rrbracket, \llbracket w^T X_- \rrbracket).$$

Dually, for an arbitrary $w \neq 0$ by putting $v = (A^{-1})^T w$, we get

$$D_{CS}(\llbracket w^T X_+ \rrbracket, \llbracket w^T X_- \rrbracket) = D_{CS}(\llbracket v^T (AX_+) \rrbracket, \llbracket v^T (AX_-) \rrbracket).$$

The assertion of the theorem follows directly from the above inequalities and (5). \square

It is easy to notice that analogously one can show that D_{CS} is translation invariant.

3.2 Data with Gaussian distribution

We proceed to the case when the data was generated from the normal distribution. Although in practice the datasets are discrete, we perform the calculations on the original continuous distributions (in next section we obtain approximation of the densities which were used to generate the data by gaussian kernel density estimation).

Let us recall that the multivariate normal density $\mathcal{N}(m, \Sigma)$ in \mathbb{R}^d with mean m and covariance matrix Σ is given by

$$\mathcal{N}(m, \Sigma)(x) = \frac{1}{(2\pi)^{d/2} (\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2} \|x - m\|_{\Sigma}^2\right),$$

where $\|\cdot\|_{\Sigma}$ denotes the Mahalanobis norm given by $\|x\|_{\Sigma}^2 = x^T \Sigma^{-1} x$.

In our considerations we will use the following well-known [20] formula for the scalar product of two normal densities:

$$\int \mathcal{N}(m_1, \Sigma_1) \mathcal{N}(m_2, \Sigma_2) = \mathcal{N}(m_1 - m_2, \Sigma_1 + \Sigma_2)(0). \quad (6)$$

Observe that from the above we easily conclude the value of the Renyi's quadratic entropy of the normal density:

$$\begin{aligned} H_2(\mathcal{N}(m, \Sigma)) &= -\log(\mathcal{N}(0, 2\Sigma)(0)) \\ &= \frac{d}{2} \log(4\pi) + \frac{1}{2} \log \det \Sigma. \end{aligned} \quad (7)$$

Theorem 3.2. *Let us consider the data X which was generated from the normal density $\mathcal{N}(m, \Sigma)$. Then the function*

$$S \ni v \rightarrow H_2(v^T X)$$

attains maximum for v being the eigenvector corresponding to the maximal eigenvalue of Σ .

Proof. One can easily check that since X has the density $\mathcal{N}(m, \Sigma)$, the projection $v^T X$ of X onto \mathbb{R} has the density

$$\mathcal{N}(v^T m, v^T \Sigma v),$$

and therefore by (7)

$$H_2(v^T X) = \frac{1}{2} \log(4\pi) + \frac{1}{2} \log(v^T \Sigma v) \text{ for } v \in S.$$

Consequently to maximize the Renyi's entropy we have to maximize the value of $v^T \Sigma v$. To do so, let us take as the base of \mathbb{R}^d the orthonormal vectors f_1, \dots, f_d which diagonalize Σ , ordered decreasingly according to the eigenvalues $\lambda_1 \geq \dots \geq \lambda_d$ of Σ . Then clearly

$$v^T \Sigma v = \sum_{i=1}^d \lambda_i v_i^2, \quad (8)$$

where v has the coefficients v_1, \dots, v_d in the considered base. Now one can easily verify by applying Lagrange multipliers that (8) under the condition $\|v\|^2 = v_1^2 + \dots + v_d^2 = 1$ is maximized for v which has coefficients $1, 0, \dots, 0$ in the base f_1, \dots, f_d , which means exactly that the maximum is attained for $v = f_1$. \square

Observe that the above result says that the information is minimal when the projection is such that the resulting density has the smallest possible variance (or in other words when it is maximally concentrated).

To present intuition concerning the Cauchy-Schwarz divergence and information potential we will consider the case of two classes with covariances proportional to identity. The result says that the crucial in this case is the projection onto line going through the means of both groups. Observe that this coincides with our intuition concerning the discrimination of those groups. Moreover, as we show in the next section, an analogous result holds for the limiting case of arbitrary sets.

Let us recall that if the data X was generated according to the distribution $\mathcal{N}(m, \Sigma)$, then $v^T X$ comes from the distribution $\mathcal{N}(v^T m, v^T \Sigma v)$. Consequently, if $\Sigma = \alpha I$ and $\|v\| = 1$ then $v^T X$ has the distribution $\mathcal{N}(v^T m, \alpha)$.

Theorem 3.3. *Let X_+, X_- be data generated by the normal densities $\mathcal{N}(m_+, \alpha_+ I)$ and $\mathcal{N}(m_-, \alpha_- I)$ with different means $m_+ \neq m_-$. Then the maximum of*

$$S \ni v \rightarrow D_{CS}(\mathcal{N}(v^T m_+, \alpha_+), \mathcal{N}(v^T m_-, \alpha_-))$$

and simultaneously minimum of

$$S \ni v \rightarrow \dot{\mathcal{P}}^*(\mathcal{N}(v^T m_+, \alpha_+), \mathcal{N}(v^T m_-, \alpha_-))$$

is attained for

$$v = \pm \frac{m_+ - m_-}{\|m_+ - m_-\|} \quad (9)$$

Proof. Since the covariances of X_+ and X_- equal $\alpha_{\pm}I$, the values of $H_2(\mathcal{N}(v^T m_+, \alpha_+))$ and $H_2(\mathcal{N}(v^T m_-, \alpha_-))$ do not depend on $v \in S$. This means the maximization of $S \ni v \rightarrow D_{CS}(\mathcal{N}(v^T m_+, \alpha_+), \mathcal{N}(v^T m_-, \alpha_-))$ is equivalent to minimization of $\dot{p}^x(\mathcal{N}(v^T m_+, \alpha_+), \mathcal{N}(v^T m_-, \alpha_-))$.

Consequently, we arrive at the problem of finding minimum of

$$\begin{aligned} S \ni v &\rightarrow \int \mathcal{N}(v^T m_+, \alpha_+) \mathcal{N}(v^T m_-, \alpha_-) \\ &= \frac{1}{\sqrt{2\pi(\alpha_+ + \alpha_-)}} \exp\left(-\frac{1}{2(\alpha_+ + \alpha_-)} \|v^T(m_+ - m_-)\|^2\right), \end{aligned}$$

which is equivalent to the search of maximum of

$$S \ni v \rightarrow \|v^T(m_+ - m_-)\|^2.$$

By the Cauchy-Schwarz inequality we trivially obtain that the above function attains its maximum for

$$v = \pm \frac{m_+ - m_-}{\|m_+ - m_-\|}$$

□

Let us interpret the Theorem 3.3 from the discrimination point of view. If we know that data from each class comes from the normal densities $\mathcal{N}(m_{\pm}, \alpha_{\pm}I)$, then the optimal projection from the information point of view is onto the line spanned by v given by (9).

4 Theory: largest margin classifiers

The core idea behind the Support Vector Machine model is to construct a linear classifier which maximizes the margin between the closest samples of opposite classes. In the simplest, linearly separable case, these closest points are referred to as support vectors. It is easy to see, that if we fix the value of support vector projections on v to $+1/-1$ then the margin M can be expressed as $1/\|v\|$ which leads to the following optimization problem.

Optimization problem: Largest margin linear classifier

$$\begin{aligned} \underset{v, b}{\text{maximize}} \quad & M = \frac{1}{\|v\|} \\ \text{subject to} \quad & y_i(v^T x_i - b) \geq 1, \quad i = 1, \dots, N \end{aligned}$$

The above problem can be reformulated in terms of minimal distance $d(v^T X_+, v^T X_-)$ between classes projections on unit length vector v , where

$$d(P_+, P_-) := \min\{|p_+ - p_-| : p_+ \in P_+, p_- \in P_-\}.$$

Reformulation: Largest margin linear classifier

$$\begin{aligned}
& \underset{v, b}{\text{maximize}} && M = d(v^T X_+, v^T X_-) \\
& \text{subject to} && \text{sign}(v^T x_i - b) = y_i, \quad i = 1, \dots, N \\
& && \|v\| = 1
\end{aligned}$$

One of the possible generalizations of this concept lies in building a multi-threshold linear classifier and maximizing all the resulting thresholds (in particular – to maximize the smallest of the thresholds). Figure 3 shows that such model can increase the size of the resulting margin even in case of the very simple dataset consisting of four points in \mathbb{R}^2 .

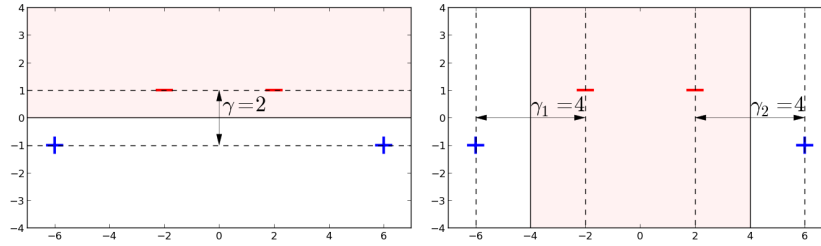


Figure 3: Example of the large margin multithreshold linear classifier (on the right) obtaining bigger margin than large margin linear classifier (on the left) on the simple dataset

From the optimization perspective the only required modification is removal of the linear separation constraint.

Optimization problem: Largest margin multithreshold linear classifier

$$\begin{aligned}
& \underset{v}{\text{maximize}} && M = d(v^T X_+, v^T X_-) \\
& \text{subject to} && \|v\| = 1
\end{aligned}$$

Such formulation can lead to arbitrary number of resulting thresholds, for example if we consider a dataset consisting of points $x_i = i, N$, $y_i = (-1)^i$, $i = 1, \dots, N$, the resulting optimal classifier would have $N - 1$ thresholds of form $t_i = i + 0.5$ (for $v = 1$). If we limit the number of resulting thresholds to k (to remove the risk of overfitting) then we end up with k -level multithreshold linear classifier.

Optimization problem: Largest margin k -level multithreshold linear classifier

$$\begin{aligned}
& \underset{v}{\text{maximize}} && M = d(v^T X_+, v^T X_-) \\
& \text{subject to} && -\infty = t_0 < \dots < t_i < \dots < t_{k+1} = \infty \\
& && v^T X_+ \subset \bigcup_{1 \leq i \leq k: 2|i} (t_i, t_{i+1}), \\
& && v^T X_- \subset \bigcup_{1 \leq i \leq k: 2|i} (t_{i-1}, t_i), \\
& && \|v\| = 1
\end{aligned}$$

It is easy to see that for $k = 1$ the above problem reduces to the SVM problem which can be solved in the polynomial time. However it appears that even in case of fixed $k = 2$ the resulting decision problem is probably NP-hard [11]. In the following subsection we will introduce method of construction of such large margin multithreshold classifier which will aim at the maximization of the margins while in the same time trying to reduce the amount of thresholds.

4.1 Preliminaries

A common method of density estimation is the kernel density estimation [6]. In the general case of data P in \mathbb{R} , we typically choose a parameter $\sigma > 0$ (often called window width), and approximate the density of the underlying distribution by

$$\llbracket P \rrbracket_\sigma := \frac{1}{|P|} \sum_{p \in P} \mathcal{N}(p, \sigma^2).$$

Although there are formulas for the optimal choice of σ when the data comes from the normal distribution, in general the optimal choice of σ is a nontrivial task which can hardly be automatized. Intuitively, for large σ the obtained density tends to become one large Gaussian, while for sufficiently small we obtain almost atom measures at each element of X . In the first case we lose important information about the local properties of the data, while the second typically leads to overfitting.

To present intuition we study the limiting cases. At first we consider the limiting case with $\sigma \rightarrow 0$, where we show that if in the linearly separable case we start from v which linearly separates the data for the case of cross-information potential, we arrive at the largest margin problem (whose solution is given by SVM). This motivates the procedure which we often apply in the next section of starting the optimization from the SVM solution. However, the minimization of cross-information potential potentially leads to overfitting where every point is memorized (although we still maximize the possible margins). Next we formally show that for $\sigma \rightarrow \infty$ our data behaves (from the point of D_{CS}) as two large Gaussians – we arrive at the formula which is an analogue of the results from the previous section. As a result we have a strong regularizing term which prevents creation of too many thresholds⁵. This supports the thesis that in the classifi-

⁵Creation of large number of thresholds often leads to overfitting

cation we should consider the whole Cauchy-Schwarz divergence, as it contains the cross-information potential term, which aims to maximize the margin, regularized by the sum of the Renyi's quadratic entropies of both classes. This theoretical observation is supported in the next section by empirical evaluation (see also Figure 4), which shows that the single use of cross-information potential in classification often leads to the unnecessary high number of thresholds, which has consequences in suboptimal classification results.

4.2 Margins maximization

In this section we are going to show that minimization of the cross information potential with small window size $\sigma \rightarrow 0$ leads to maximization of the margin width between the classes. Simultaneously, this will imply the existence of many local minima's.

We begin with the following proposition.

Proposition 4.1. *Let $P_+, P_- \subset \mathbb{R}$ be given, $\sigma > 0$. Then*

$$\dot{\psi}^\times(\llbracket P_+ \rrbracket_\sigma, \llbracket P_- \rrbracket_\sigma) \leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d^2(P_+, P_-)}{2\sigma^2}\right), \quad (10)$$

$$\dot{\psi}^\times(\llbracket P_+ \rrbracket_\sigma, \llbracket P_- \rrbracket_\sigma) \geq \frac{1}{\sqrt{2\pi}\sigma|P_+| \cdot |P_-|} \exp\left(-\frac{d^2(P_+, P_-)}{2\sigma^2}\right). \quad (11)$$

Proof. Let us choose $\bar{p}_+ \in P_+$ and $\bar{p}_- \in P_-$ such that $|\bar{p}_+ - \bar{p}_-| = D = d(P_+, P_-)$, then

$$\int \frac{1}{|P_+|} \mathcal{N}(p_+, \sigma^2) \cdot \frac{1}{|P_-|} \mathcal{N}(p_-, \sigma^2) = \frac{1}{|P_+| \cdot |P_-|} \frac{\exp(-D^2/2\sigma^2)}{\sqrt{2\pi}\sigma}.$$

On the other hand

$$\begin{aligned} & \int \sum_{p_+ \in P_+} \frac{1}{|P_+|} \mathcal{N}(\bar{p}_+, \sigma^2) \cdot \sum_{p_- \in P_-} \frac{1}{|P_-|} \mathcal{N}(\bar{p}_-, \sigma^2) \\ & \leq \sum_{p_+ \in P_+, p_- \in P_-} \frac{1}{|P_+| \cdot |P_-|} \frac{\exp(-D^2/2\sigma^2)}{\sqrt{2\pi}\sigma} = \frac{\exp(-D^2/2\sigma^2)}{\sqrt{2\pi}\sigma}. \end{aligned}$$

□

Given $v \neq 0$ and $\sigma > 0$ we put

$$\dot{\psi}_\sigma^\times(v) := \dot{\psi}^\times(\llbracket v^T X_+ \rrbracket_\sigma, \llbracket v^T X_- \rrbracket_\sigma).$$

We say that $v \in S$ *linearly separates* X_- from X_+ if

$$\inf(v^T X_+) \geq \sup(v^T X_-).$$

First, we show that if we start the gradient descent method of $\dot{\psi}_\sigma^\times(\cdot)$ with sufficiently small $\sigma > 0$ from the vector which linearly separates the classes, we will remain in the set of vectors which discriminate the classes.

Theorem 4.1. *We assume that $v \in S$ linearly separates X_- from X_+ and that $\sigma > 0$ is such that*

$$\sigma < \frac{d(v^T X_+, v^T X_-)}{\sqrt{2 \log(|X_+| \cdot |X_-|)}}.$$

Then steepest descent for minimization of $\dot{q}_\sigma^\times(\cdot)$ leads to the choice of v' which also linearly separates X_- from X_+ .

Proof. Let $v : [0, \bar{t}] \rightarrow S$, $v(0) = v$, $v(\bar{t}) = v'$ be an arbitrary continuous curve (in particular given by the steepest descent method) along which the value of $t \rightarrow \dot{q}_\sigma^\times(v(t))$ does not increase.

Suppose that the assertion does not hold. This means that there exists $x_+ \in X_+$, $x_- \in X_-$ and $t \in [0, \bar{t}]$ such that

$$v(t)^T x_+ \leq v(t)^T x_-.$$

By the continuity we conclude that there exists $t_0 \leq t$ such that

$$v(t_0)^T x_+ = v(t_0)^T x_-.$$

This means that $d(v(t_0)^T X_+, v(t_0)^T X_-) = 0$, and consequently by the previous proposition we get

$$\dot{q}_\sigma^\times(v(t_0)) \geq \frac{1}{\sqrt{2\pi\sigma}|X_+| \cdot |X_-|}.$$

But from the assumptions we know that the cross-information potential does not increase with t , which means that

$$\dot{q}_\sigma^\times(v(t_0)) \leq \dot{q}_\sigma^\times(v(0)) \leq \frac{\exp(-d^2(v^T X_+, v^T X_-)/(2\sigma^2))}{\sqrt{2\pi\sigma}}.$$

Joining this with the previous inequality, after obvious calculations, we obtain

$$d(v^T X_+, v^T X_-) \leq \sqrt{2\sigma \log^{1/2}(|X_+| \cdot |X_-|)},$$

a contradiction. □

Suppose that X_- is linearly separable from X_+ . Let

$$S_{LS}(X_+, X_-) := \{v \in S : v \text{ linearly separates } X_- \text{ from } X_+\}.$$

Let $M_{SVM}(X_-, X_+)$ denote the maximal possible margin along v which linearly separates X_+ from X_- :

$$M_{SVM}(X_-, X_+) := \sup\{d(v^T X_+, v^T X_-) : v \in S_{LS}(X_+, X_-)\}.$$

By v_{SVM} we denote the unique element which realizes the above maximum. Clearly, it is given by the normalized solution to the SVM process.

Now we are going to show that in the limiting case $\sigma \rightarrow 0$, we converge to the solution of the SVM procedure – in other words we aim at the maximization of the linear margin between classes.

Theorem 4.2. Consider linearly separable classes X_- and X_+ . Let $\bar{v} \in S_{LS}(X_+, X_-)$ denote an arbitrary point which realizes the minimum of the cross-information potential:

$$\dot{\psi}_\sigma^\times(\bar{v}) = \min\{\dot{\psi}_\sigma^\times(v) : v \in S_{LS}(X_+, X_-)\}.$$

Then the resulting linear classifier's margin $d(\bar{v}^T X_+, \bar{v}^T X_-)$ is at least as big as

$$M_{SVM}(X_-, X_+) - \sigma\sqrt{2\log(|X_+| \cdot |X_-|)}.$$

Proof. Let us first estimate the value of $\dot{\psi}_\sigma^\times(v_{SVM})$ by applying Proposition 4.1 with $P_\pm = \bar{v}^T X_\pm$

$$\dot{\psi}_\sigma^\times(v_{SVM}) \leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{M(X_-, X_+)^2}{2\sigma^2}\right).$$

So let us now choose an arbitrary \bar{v} which separates X_+ from X_- which realizes the minimum of the cross-information potential. Then

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{M(X_-, X_+)^2}{2\sigma^2}\right) &\geq \dot{\psi}_\sigma^\times(v_{SVM}) \geq \dot{\psi}_\sigma^\times(\bar{v}) \\ &\geq \frac{1}{\sqrt{2\pi}\sigma|X_+||X_-|} \exp\left(-\frac{M(\bar{v}; X_+, X_-)^2}{2\sigma^2}\right), \end{aligned}$$

which directly yields the assertion of the theorem. \square

Now we discuss the minimization of cross-information potential over the whole S . It occurs that in the limiting case $\sigma \rightarrow 0$ it results in the margin maximization. Let $M(X_-, X_+)$ denote the maximal possible margin along $v \in S$

$$M(X_-, X_+) := \sup\{d(v^T X_+, v^T X_-) : v \in S\}.$$

By applying similar reasoning as in the proof of Theorem 4.2 we obtain that the minimization of cross-information potential leads to the maximization of multiple margins in the multithreshold classifier.

Theorem 4.3. Consider classes X_- and X_+ . Let $\bar{v} \in S$ denote an arbitrary point which realizes the minimum of cross-information potential:

$$\dot{\psi}_\sigma^\times(\bar{v}) = \min\{\dot{\psi}_\sigma^\times(v) : v \in S\}.$$

Then the resulting multithreshold linear classifier's margins $d(\bar{v}^T X_+, \bar{v}^T X_-)$ are at least as big as

$$M(X_-, X_+) - \sigma\sqrt{2\log(|X_+| \cdot |X_-|)}.$$

The above theorem leads to the conclusion that cross-information potential (without additional regularizing terms) leads to the construction of largest margin multithreshold classifier. However, it lacks the ability to control the number of resulting thresholds and as a result, for sufficiently small σ , it may construct an interval for each data point, which leads to overfitting. A real life example of sonar dataset from UCI repository is given in Figure 4, which shows the comparison of minimization of $\dot{\psi}^\times$ versus maximization of Cauchy-Schwarz divergence. Following section shows how introduction of the classes' entropies to the optimization process causes reduction of the model's complexity (in a limiting case to the linear classifier).

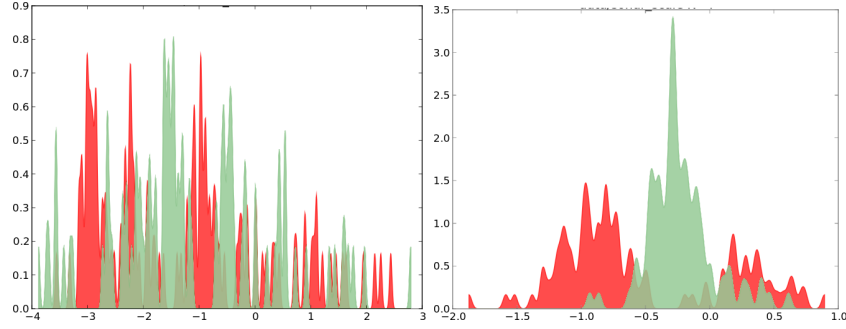


Figure 4: Sample kernel density estimation of projected sonar dataset with small σ using \dot{p}^* optimization (on the left) and D_{CS} (on the right).

4.3 Regularization

We show that the analogue of the Theorem 3.3 holds also for the limiting case when we increase the window width to infinity. This will result in construction of the linear classifier (limiting reduction of the number of thresholds).

We consider only the case when the window width σ is set to be equal for both classes. We recall that for $P \subset \mathbb{R}$ and $\sigma > 0$ we put

$$\llbracket P \rrbracket_\sigma := \frac{1}{|P|} \sum_{p \in P} \mathcal{N}(p, \sigma^2).$$

Thus $\llbracket P \rrbracket_\sigma$ denotes the kernel density estimation based on the set P with window width σ . We begin with the following observation.

Proposition 4.2. *Let $P_+, P_- \subset \mathbb{R}$ be given, and let the window width $\sigma > 0$ be fixed. Then*

$$D_{CS}(\llbracket P_+ \rrbracket_\sigma, \llbracket P_- \rrbracket_\sigma) = \frac{1}{2\sigma^2} (m_+ - m_-)^2 + \mathcal{O}(\sigma^{-4}) \text{ as } \sigma \rightarrow \infty, \quad (12)$$

where m_\pm denote the means of P_\pm .

Proof. We denote elements of P_+ by p_+ , and elements of P_- by p_- .

We have

$$\begin{aligned} \int \llbracket P_+ \rrbracket_\sigma \llbracket P_- \rrbracket_\sigma &= \frac{1}{|P_+||P_-|} \sum_{p_+, p_-} \mathcal{N}(p_+ - p_-, 2\sigma^2) \\ &= \frac{1}{\sqrt{4\pi\sigma^2}|P_+||P_-|} \sum_{p_+, p_-} \exp(-(p_+ - p_-)^2/(4\sigma^2)). \end{aligned}$$

Since $\exp(h) = 1 + h + \mathcal{O}(h^2)$ and $\log(1 + h) = h + \mathcal{O}(h^2)$ for small h , the above

equality implies that for large σ

$$\begin{aligned}
& \log \int \llbracket P_+ \rrbracket_\sigma \llbracket P_- \rrbracket_\sigma \\
&= -\log(2\sigma\sqrt{\pi}) \\
&\quad + \log\left(1 - \frac{1}{4\|P_+\|P_-\|\sigma^2}\sum_{p_+,p_-}(p_+ - p_-)^2\right) + \mathcal{O}(\sigma^{-4}) \\
&= -\log(2\sigma\sqrt{\pi}) - \frac{1}{4\|P_+\|P_-\|\sigma^2}\sum_{p_+,p_-}(p_+ - p_-)^2 + \mathcal{O}(\sigma^{-4}).
\end{aligned}$$

Consequently

$$\begin{aligned}
& D_{CS}(\llbracket P_+ \rrbracket_\sigma, \llbracket P_- \rrbracket_\sigma) \\
&= \log \int \llbracket P_+ \rrbracket_\sigma^2 + \log \int \llbracket P_- \rrbracket_\sigma^2 - 2 \log \int \llbracket P_+ \rrbracket_\sigma \llbracket P_- \rrbracket_\sigma \\
&= \frac{1}{4\sigma^2} \left(-\frac{1}{\|P_+\|^2} \sum_{p_+,p'_+}(p_+ - p'_+)^2 - \frac{1}{\|P_-\|^2} \sum_{p_-,p'_-}(p_- - p'_-)^2 \right. \\
&\quad \left. + \frac{2}{\|P_+\|P_-\|} \sum_{p_+,p_-}(p_+ - p_-)^2 \right) + \mathcal{O}(\sigma^{-4}).
\end{aligned}$$

By applying obvious calculations we obtain that

$$\begin{aligned}
& -\frac{1}{\|P_+\|^2} \sum_{p_+,p'_+}(p_+ - p'_+)^2 - \frac{1}{\|P_-\|^2} \sum_{p_-,p'_-}(p_- - p'_-)^2 \\
&\quad + \frac{2}{\|P_+\|P_-\|} \sum_{p_+,p_-}(p_+ - p_-)^2 \\
&= -2\left(\frac{1}{\|P_+\|} \sum_{p_+} p_+^2 - \left(\frac{1}{\|P_+\|} \sum_{p_+} p_+\right)^2\right) \\
&\quad - 2\left(\frac{1}{\|P_-\|} \sum_{p_-} p_-^2 - \left(\frac{1}{\|P_-\|} \sum_{p_-} p_-\right)^2\right) \\
&\quad + \frac{2}{\|P_+\|} \sum_{p_+} p_+^2 + \frac{2}{\|P_-\|} \sum_{p_-} p_-^2 - \frac{4}{\|P_+\|P_-\|} \sum_{p_+} p_+ \sum_{p_-} p_- \\
&= -4m_+m_- + 2m_+^2 + 2m_-^2 = 2(m_+ - m_-)^2.
\end{aligned}$$

□

Observe that the constant C in $\mathcal{O}(\sigma^{-4}) = C\sigma^{-4}$ in the (12) can be estimated from the proof by an increasing function of $\sum_{p_+} |p_+|^2 + \sum_{p_-} |p_-|^2$.

Theorem 4.4. *We consider classes X_+ and X_- . We assume additionally that class centers $m_\pm \in \mathbb{R}^d$ are such that $m_+ \neq m_-$, where m_\pm denote the means of X_\pm . We put*

$$v_\infty = \frac{m_+ - m_-}{\|m_+ - m_-\|}.$$

For $\sigma > 0$ let $v_\sigma \in S$ denote the argument for which the function

$$D_{CS}^\sigma(v) : S \ni v \rightarrow D_{CS}(\llbracket v^T X_+ \rrbracket_\sigma, \llbracket v^T X_- \rrbracket_\sigma)$$

takes the maximum value. Then v_σ tends to $\pm v_\infty$ with increasing σ , that is

$$\min(\|v_\sigma - v_\infty\|, \|v_\sigma + v_\infty\|) = \mathcal{O}(\sigma^{-1}) \text{ as } \sigma \rightarrow \infty.$$

Proof. Clearly, by the Proposition 4.2

$$\frac{2\sigma^2}{\|m_+ - m_-\|^2} D_{CS}^\sigma(v) = \langle v, v_\infty \rangle^2 + \mathcal{O}(\sigma^{-2}), \quad (13)$$

and the constant in $\mathcal{O}(\sigma^{-2})$ can be bounded by an increasing function of $\sum_{x_+} \|x_+\|^2 + \sum_{x_-} \|x_-\|^2$.

Consider $v_\sigma \in S$. Without loss of generality by taking $-v_\sigma$ in place of v_σ , if necessary, and applying the fact that D_{CS}^σ is an even function, we may assume that v_σ is nearer to v_∞ than to $-v_\infty$, that is $\|v_\sigma - v_\infty\| \leq \|v_\sigma + v_\infty\|$. We are going to estimate from above the value of $\|v_\sigma - v_\infty\|$. Observe first that

$$\begin{aligned} \langle v_\sigma, v_\infty \rangle &= \frac{1}{2}(\|v_\sigma\|^2 + \|v_\infty\|^2 - \|v_\sigma - v_\infty\|^2) \\ &= 1 - \frac{1}{2}\|v_\sigma - v_\infty\|^2. \end{aligned} \quad (14)$$

This trivially yields that $\langle v_\sigma, v_\infty \rangle \geq 0$.

On the other hand, since D_{CS}^σ takes maximum in v_σ , by applying (13) twice, we get

$$\begin{aligned} \langle v_\sigma, v_\infty \rangle^2 &\geq \frac{2\sigma^2}{\|m_+ - m_-\|^2} D_{CS}^\sigma(v_\sigma) - C'\sigma^2 \\ &\geq \frac{2\sigma^2}{\|m_+ - m_-\|^2} D_{CS}^\sigma(v_\infty) - C'\sigma^2 \\ &\geq \langle v_\infty, v_\infty \rangle^2 - C''\sigma^{-2} = 1 - C''\sigma^{-2} \end{aligned}$$

for certain $C', C'' > 0$. Since $\sqrt{1-h} \geq 1-h$ (for $h \geq 0$) and $\langle v_\sigma, v_\infty \rangle$ is nonnegative, this yields that

$$\langle v_\sigma, v_\infty \rangle \geq \sqrt{1 - C''\sigma^{-2}} \geq 1 - C''\sigma^{-2}.$$

By applying (14) we conclude that $\|v_\sigma - v_\infty\|^2 < 2C''\sigma^{-2}$, which yields

$$\|v_\sigma - v_\infty\| = \mathcal{O}(\sigma^{-1}).$$

□

4.4 Classification theory

Let us recall that the objective function

$$D_{CS}(f, g) = -(H_2(f) + H_2(g) + 2 \log \dot{p}^\times(f, g)),$$

consists of two parts, the entropy term $H_2(f) + H_2(g)$ which serves the regularization purpose and $\dot{p}^\times(f, g)$ which ensures optimal discrimination of the classes. Maximization of D_{CS} and classifying data based on the 1-dimensional kernel density estimation leads to the construction of multithreshold linear classifier. Optimization procedure tries to simultaneously maximize the margins between classes and to minimize the number of resulting thresholds.

As Anthony [7] showed, the considered class of classifiers have bounded generalization error dependent on the number of thresholds k :

Theorem 4.5. *Generalization bounds (Anthony, 2004 [7]) With probability at least $1 - \delta$, for N points in \mathbb{R}^d :*

$$E \leq E_{emp} + \sqrt{\frac{8}{N} \left((d+k-1) \log \left(\frac{2eNk}{d+k-1} \right) + \log \left(\frac{14k^2}{\delta} \right) \right)}$$

where E is the generalization error and E_{emp} is the training (empirical) error.

As it has been previously shown, minimization of the Renyi's entropy leads to the choice of projections where each class is as condensed as possible. In a natural way this means that this process leads to the minimization of number of resulting thresholds (the value of estimated density is monotonically decreasing when we move away from the closest point with Gaussian function centered in it).

The following theorem shows that for k -level threshold linear classifier restricted to the sphere, the generalization bounds can be improved by maximizing the margin M .

Theorem 4.6. *Generalization bounds with margin (Anthony, 2004 [7]) With probability at least $1 - \delta$, for N points in \mathbb{R}^d such that $\|x_i\| \leq 1$, $\|v\| = 1$ and margin $M \in (0, 1]$:*

$$E \leq E_{emp} + \sqrt{\frac{8}{N} \left(\frac{1152}{M^2} \log(9N) + k \log\left(\frac{10}{M}\right) + \log\left(\frac{4}{\delta}\right) \right)}$$

According to Theorems 4.1 and 4.2 minimization of \dot{p}^x leads (in the limiting case) to the maximization of the separating margins. So our method is truly aimed at structural risk minimization. We search for such multithreshold linear classifier which minimizes the generalization error through selection of the structurally simplest hypothesis. This shows another similarity to the SVM model, but adapted to multithreshold case.

5 Practical considerations

In this section we deal with some practical considerations regarding our optimization problem, which lies in maximizing the Cauchy-Schwarz divergence of the kernel density estimation of projections of our data. As it has been proven in Theorem 3.1, this problem is scale invariant, so we can constrain domain of searched parameters into the unit sphere in \mathbb{R}^d . In practice this limitation reduces not only the parameter space, but also the risk of numerical instability, while coming at no additional computational cost.

In the optimization we apply the typical steepest ascent approach. It is a common knowledge that such procedure can be performed for maximization of function f on the sphere by simply projecting the gradient onto the tangent hyperplane and performing the usual line search procedure on the big circle given by gradient's direction. Given the starting point v_0 such that $\|v_0\| = 1$ it can be expressed as following iterative procedure for step sizes α_t :

$$\begin{aligned} h_t &= \nabla_v f(v_t) - \langle \nabla_v f(v_t), v_t \rangle v_t, \\ v_{t+1} &= v_t \cos(\alpha_t) + \sin(\alpha_t) h_t / \|h_t\|. \end{aligned}$$

Let us now summarize the problem of D_{CS} maximization with Silverman's rule for kernel density estimator window width.

MELC optimization problem. *Given sets $X_+, X_- \subset \mathbb{R}^d$*

$$\begin{aligned}
& \underset{v \in \mathbb{R}^d}{\text{maximize}} \quad - (H_2^+(v) + H_2^-(v) + 2 \log \dot{\psi}_{X_+ X_-}^\times(v)) \\
& \text{subject to} \quad \|v\| = 1 \\
& \text{where} \\
& H_2^\pm(v) = -\log \dot{\psi}_{X_\pm X_\pm}^\times(v) \\
& \dot{\psi}_{AB}^\times(v) = \frac{1}{\sqrt{2\pi V_{AB}(v)} \cdot |A||B|} \sum_{a \in A, b \in B} \exp\left(-\frac{\langle v, a-b \rangle^2}{2V_{AB}(v)}\right), \\
& V_{AB}(v) = V_A(v) + V_B(v), \\
& V_A(v) = ((4/3)^{1/5} |A|^{-1/5} \sigma_{v, T_A})^2.
\end{aligned}$$

In order to perform steepest ascent optimization we need to compute gradient of D_{CS} function. We present its final formula and omit its obvious derivation.

$$\begin{aligned}
\nabla D_{CS}(v) &= \frac{\nabla \dot{\psi}_{X_+ X_+}^\times(v)}{\dot{\psi}_{X_+ X_+}^\times(v)} + \frac{\nabla \dot{\psi}_{X_- X_-}^\times(v)}{\dot{\psi}_{X_- X_-}^\times(v)} - \frac{2 \nabla \dot{\psi}_{X_+ X_-}^\times(v)}{\dot{\psi}_{X_+ X_-}^\times(v)}, \\
\nabla \dot{\psi}_{AB}^\times(v) &= \frac{1}{2V_{AB}(v) \sqrt{2\pi V_{AB}(v)} \cdot |A||B|} \sum_{a \in A, b \in B} \exp\left(-\frac{\langle v, a-b \rangle^2}{2V_{AB}(v)}\right) \\
&\quad \left\{ \left(\frac{\langle v, a-b \rangle^2}{2V_{AB}(v)} - 1 \right) \nabla V_{AB}(v) - 2\langle v, a-b \rangle (a-b) \right\}, \\
\nabla V_{AB}(v) &= \nabla V_A(v) + \nabla V_B(v), \\
\nabla V_A(v) &= \frac{\left(\frac{4}{3}\right)^{\frac{2}{5}}}{|A|^{12/5}} \left(|A| \cdot \sum_{a \in A} \langle v, a \rangle a - \sum_{a \in A} \langle v, a \rangle \cdot \sum_{a \in A} a \right).
\end{aligned}$$

It is easy to see that computation of both function value and its gradient is computationally expensive ($\mathcal{O}(N^2)$, where N is the size of the training set). This issue is partially compensated by fact that in practice it is sufficient to perform just few steps of this process in order to find a local maxima. In the basic approach we choose random starting points from the sphere, run optimizations from them and select the one yielding the biggest value. However, it is also possible to start optimization from the solution given by some computationally cheap model, like for example a perceptron or a linear SVM with $C = 1$. As a result, we can obtain a reasonable solution in quite short time (using just one optimization procedure). These methods are further investigated in the Evaluation Section.

Classifier complexity

Classification using the actual density estimators on \mathbb{R} requires $\mathcal{O}(N)$ operations (each training point has impact on the classification). This issue can be overcome by constructing the actual k -threshold linear classifier from this density by search for points t_1, \dots, t_k such that $\llbracket v^T X_+ \rrbracket(t_i) = \llbracket v^T X_- \rrbracket(t_i)$ (see algorithm in Figure 5).

```

1:  $x_1, \dots, x_N \leftarrow \text{sort}(v^T X_+ \cup v^T X_-)$ 
2:  $Q \leftarrow \llbracket v^T X_+ \rrbracket - \llbracket v^T X_- \rrbracket$ 
3:  $k \leftarrow 0$ 
4: for  $i = 2$  to  $N$  do
5:   if  $\text{sign}(Q(x_{i-1})) \neq \text{sign}(Q(x_i))$  then
6:      $k \leftarrow k + 1$ 
7:      $t_k \leftarrow \text{binsearch}_{x \in (x_{i-1}, x_i)} Q(x) = 0$ 
8:   end if
9: end for
10: return  $t_1, \dots, t_k$ 

```

Figure 5: k -threshold linear classifier construction for kernel density estimation of X_{\pm} projections on given v

As a result, classification's complexity of the new points is decreased to $\mathcal{O}(d + \log(k))$ (binary search of k midpoints on \mathbb{R}). In the Evaluation Section we also show that it is sufficient to run just few iterations of *binsearch* to build such classifier. However, such operation destroys easy access to the estimation of $P(y|x)$, as similarly to other linear models we just have thresholds. In order to obtain such probabilities (confidences) one still needs to query all the training points. If such approach is too expensive one can change used kernel to the Epanechnikov or other with finite support.

Parameterization

One can use different kernel width estimator by alternating the $V_A(v)$ term (and its gradient). In particular, in order to include the kernel window width scaling factor γ it is sufficient to replace the variance term $V_A(v)$ in the previous equations with $V_A^\gamma(v) := \gamma^2 V_A(v)$, and analogously $\nabla V_A(v)$ becomes $\nabla V_A^\gamma(v) := \gamma^2 \nabla V_A(v)$. As shown in the Evaluation section, this can be beneficial as the Silverman's rule tends to overestimate the required value [4]. Size of the γ factor plays also the role of a bias-variance tradeoff coefficient. With bigger values the optimization will lead to the very simple single-threshold models (with a limiting case proven in the Theory Section), while the small values can lead to overfitting the data. Default value of $\gamma = 1$ yields quite reasonable solutions (as showed in the Evaluation Section) but results can be improved by searching for (in most cases) smaller values. In general, regularization strength grows with γ .

It is also possible to include samples weights w_x directly in the proposed formulation. The only modification needed is to put the weighted kernel density estimator

$$\llbracket P \rrbracket_\sigma = \frac{1}{\sum_{p \in P} w_p} \sum_{p \in P} w_p \mathcal{N}(p, \sigma^2).$$

It is worth noting that this weighting works on the basis of in-class weights, it cannot be directly applied to weight the whole class. On the other hand similar concept can be used to include the known input data uncertainty measure by using different σ_x for each point.

Non-linear case

For problems requiring non-linear model one can adapt the proposed approach. Direct kernelization would lead to higher computational complexity ($\mathcal{O}(N^3)$ per iteration), but there are other possible solutions. First, one can use Nystrom’s method of kernel approximation [21] which does not require such complex operations. It is also possible to apply random projection techniques [22, 1, 23], which map the input space through some non-linear function (eg. RBF) as the preprocessing step. In particular, one can use clustering methods to seed the position of RBF function (as it is done in RBF networks [24]). This problem is, however, beyond the scope of this work and should be the topic of future research.

6 Evaluation

We evaluated our method using code written in C++ with help of `boost` [25] library. Experiments were conducted on an Intel Xeon 2.67GHz machine. In the first phase we used ten well known UCI [26] binary datasets, briefly summarized in Table 1.

dataset	d	n	$ X_+ $	$ X_- $
australian	14	690	383	307
breast cancer	9	683	444	239
diabetes	8	768	268	500
fourclass	2	862	307	555
german number	24	1000	700	300
heart	13	270	150	120
ionosphere	34	351	225	126
liver-disorders	6	345	145	200
sonar	60	208	111	97
splice	60	1000	483	517

Table 1: Summary of UCI datasets used in tests.

During the second part of the evaluation we focused on real, cheminformatics data regarding chemical compounds activity prediction for selected proteins. For different models implementations we used the `scikit-learn` [27] package, implementing the popular `libSVM` [28] library. Three evaluation metrics are used in further parts of our paper: accuracy (ACC), Matthew’s Correlation Coefficient (MCC) and weighted accuracy (WAC, also known as averaged accuracy).

6.1 Toy dataset

For better understanding of our method’s characteristic we begin evaluation with XOR like dataset, composed of 100 samples from four two-dimensional Gaussians centered in points $(-1, -1)$, $(+1, +1)$ (positive samples) and $(-1, +1)$, $(+1, -1)$ (negative ones), see Figure 6.

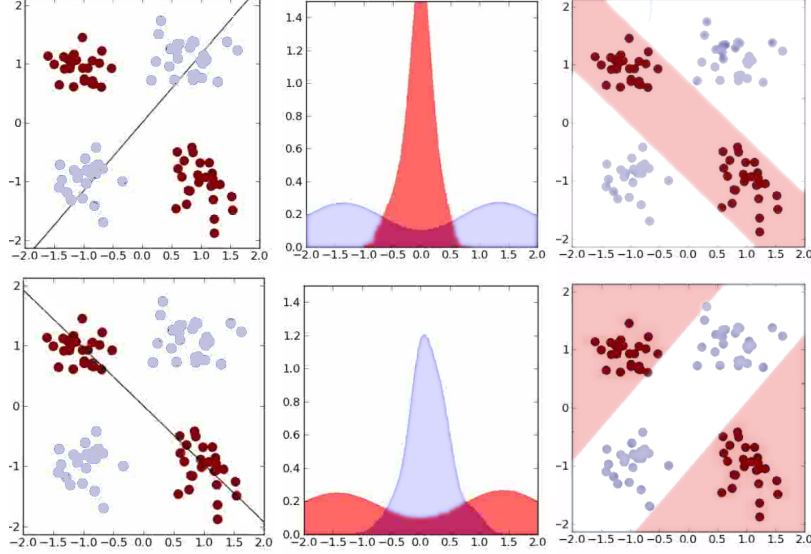


Figure 6: XOR like dataset composed of four Gaussians

Obviously this dataset is not linearly separable, but can be shattered with use of a 2-threshold linear classifier. In terms of D_{CS} this data has two (up to the center symmetry) local maxima, one around $v_1 = (\sqrt{2}/2, \sqrt{2}/2)$ and one around $v_2 = (-\sqrt{2}/2, \sqrt{2}/2)$. Solution given by v_1 has higher D_{CS} as the spiked class is much narrower (its Renyi’s entropy is lower), and as a result – smallest of the two resulting margins is bigger. One can notice, that density estimation with Silverman’s rule tends to overestimate the required kernel window size (splitted class is too flat).

Proposed method achieved almost 100% scores under all considered metrics, while the linear models (both perceptron and SVM) achieved at most 50% accuracy. Naturally, if kernelized with polynomial kernels, these methods would perform much better. This is however only a simple example to illustrate the potential benefits of multithreshold classifier while still using only the linear projection.

6.2 Impact of regularization

In the Theoretical Section we showed that minimization of \hat{p}^x leads to the separation with the large margin. However, if the chosen kernel width is too small, this may lead to overfitting issues due to the multithreshold nature of our model. In the worst case scenario, when our density estimation degenerates to almost atomic measure, we would get a perfect training set fitting with no generalization capabilities. This supports the need for the regularization based on the each classes densities' entropies and as a result optimization of D_{CS} instead of just \hat{p}^x . On the Figure 7 one can find histogram of number of thresholds in our model for UCI datasets. We use Silverman's rule for kernel width estimation, which is known to rather overestimate this value (the optimal kernel width is often smaller than the one given by Silverman). However, even in such case one can notice, that purely \hat{p}^x based optimization leads to significantly more complex models (with higher number of thresholds) .

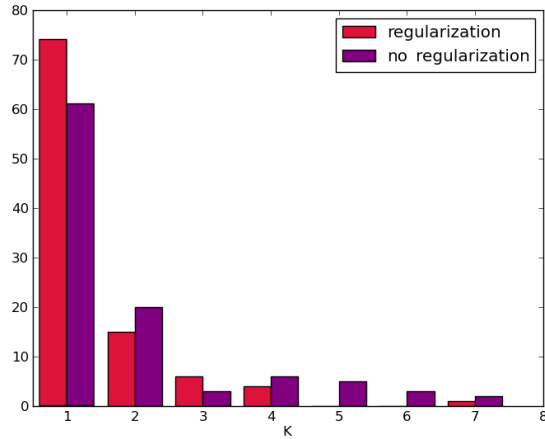


Figure 7: Histogram of number of resulting thresholds in classifiers built on the UCI datasets

6.3 D_{CS} and generalization

In the previous sections we argued that maximalization of the Cauchy-Schwarz divergence should lead to the choice of a model with good generalization capabilities. In the Figure 8 one can see how value of D_{CS} is correlated with the Matthew's Correlation Coefficient (measured on the test set) for splice dataset.

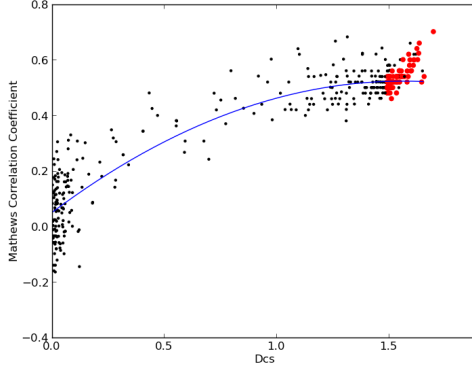


Figure 8: Correlation between D_{CS} value (on x axis) and generalization capabilities (expressed as MCC on the test sets in 10 CV) for the splice dataset.h Big dots represent local maxima of D_{CS} found during opitmization process.

Easily noticable relation suggests that D_{CS} can be truly used as a criterion for the choice of model. Pearson’s correlation coefficient between these two values for splice is about 0.9. It seems also, that it is rather resistant to the overfitting (as there is no noticable decrease in the generalization for high D_{CS} values). Correlations for the remaining datasets are reported in Table 2, all of them are statistically significant (in terms of correlation p-value).

dataset	ACC	MCC	WAC
australian	0.898	0.900	0.902
breast cancer	0.901	0.897	0.896
diabetes	0.494	0.611	0.624
fourclass	0.245	0.374	0.393
german number	0.407	0.569	0.575
heart	0.726	0.726	0.728
ionosphere	0.537	0.532	0.518
liver-disorders	0.348	0.350	0.357
sonar	0.645	0.635	0.644
splice	0.943	0.941	0.943

Table 2: Mean correlation between D_{CS} and the generaliztaion capabilities across 10-folds of cross validation.

First, we see that in all cases there is a moderate to strong positive correlation. Second, these results confirm that our method is aimed at balanced measures (like WAC and MCC) while in the same time not well suited for accuracy (which by its definition prefers non-balanced models). In further part of

our paper we focus only on these two metrics.

6.4 UCI binary classification

In the following part we will compare the efficiency of Multithreshold Entropy Linear Classifier (MELC), Support Vector Machines (SVM), Support Vector Machines with class balancing (SVM-B) and Perceptron. SVM-B is the SVM model with C value splitted into C_+ and C_- invertibly proportional to the corresponding class sizes. All experiments are performed in 10-fold cross validation.

We first investigated how well these four models work when ran with default parameters (as given in `scikit-learn` library, which means $C = 1$ for SVM models). Figure 9 shows obtained results in terms of WAC measure (results for MCC were analogous). Without tuning of any model, MELC obtained

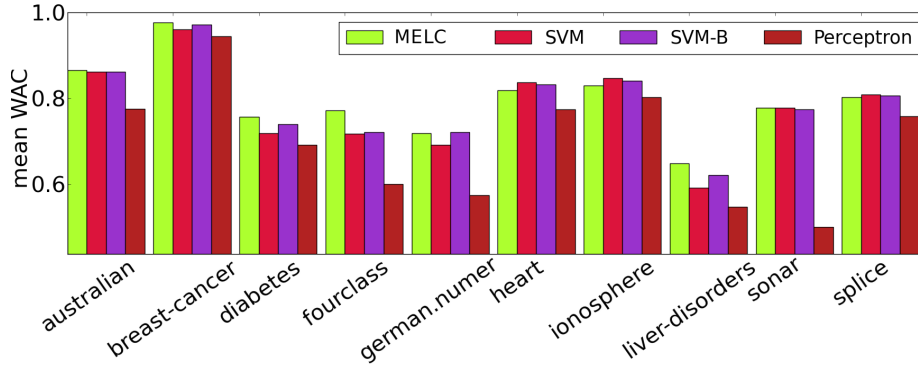


Figure 9: Comparision of 10-fold cross validation WAC scores with default parameters

results comparable with SVM for most datasets, and outperformed it for a few (including liver-disorders, fourclass and diabetes). Results of perceptron were significantly worse in all cases. In nine of ten datasets MELC build a linear classifier, and in case of fourclass dataset, a 3-threshold linear classifier. This supports our claim that the regularization prevents model from selecting too big k values. However, it is worth noting that even though MELC gained similar mean WAC as SVMs for some problems, it built different decision models. In particular, after investigation of results of individual folds, in some cases our method significantly outperformed SVM and vice-versa. It supports our claim, that even though there are important theoretical connections between these models, they result in different classifiers.

As it was previously stated, process of optimization of D_{CS} may be computationally expensive. To deal with this problem one can perform single (or few) gradient based optimization from solutions given by some other, cheaper models. Comparison of the results obtained by our approach seeded with v found

by SVMs and perceptron are plotted in Figure 10. One can notice, that such

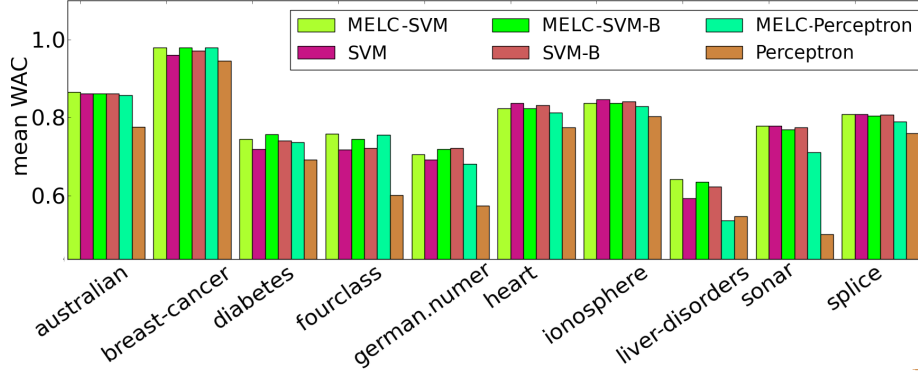


Figure 10: Comparison of 10-fold cross validation WAC scores for MELC starting from solution given by SVM, SVM-B and perceptron (with default parameters)

initialization can lead to quite reasonable solutions. Starting from perceptron solutions generally lead to much worse scores, as this model finds completely different type of solutions than MELC does. In case of SVM it seems possible to exploit already performed optimization. In particular, in our experiments rather low dimensional problems from UCI library can be well solved by starting from random points sampled uniformly from the unit sphere. In contrast, when number of dimensions is significantly higher and the optimization problem is harder it is more valuable to initialize the weights vector by running optimization from balanced SVM solution (for example, with $C = 1$). Such an approach is further used in the last section of the evaluation.

We have shown how MELC behaves when treated as non-parametric model. However, similarly to the C parameter in SVM formulation, we can control the strength of the regularization. In Table 3 one can find WAC scores for MELC (with fitted γ) as compared to SVM and balanced SVM (with fitted C). Obtained results resemble ones from the previous experiments, MELC obtained similar results to the SVM, with some datasets showing superiority of the entropy based approach. In particular, in case of fourclass dataset one can see even bigger advantage of using multithreshold function over the simple linear classifier. It is also worth noting, that MELC parameter has much more clear geometrical interpretation than SVM's parameter C (which can be seen either as abstract weight of training errors or as an upper bound on the size of Lagrange multipliers). The parameter γ , or in general the formula for $V_A(v)$, gives the estimation of optimal kernel width in one dimensional projection of A on v . There are many existing studies [29, 30, 31] and formulas for such objects, in particular it is possible to perform adaptive kernel width [32] where each point x have its own kernel width σ_x .

dataset	MELC	SVM	SVM-B
australian	0.868 [1.0]	0.862	0.862
breast cancer	0.979 [1.0]	0.969	0.972
diabetes	0.758 [1.0]	0.727	0.747
fourclass	0.843 [4.0]	0.720	0.727
german number	0.726 [1.1]	0.691	0.722
heart	0.836 [1.0]	0.837	0.838
ionosphere	0.848 [1.0]	0.862	0.860
liver-disorders	0.658 [2.9]	0.677	0.659
sonar	0.791 [1.0]	0.790	0.790
splice	0.810 [1.0]	0.810	0.810

Table 3: Comparison of 10-fold cross validation WAC scores for MELC and SVM, SVM-balanced (SVM-B) with optimized parameters. Mean number of thresholds for MELC is reported in square brackets

6.5 Compounds activity prediction

Final part of our evaluation was performed on cheminformatical data. The task is to predict whether a chemical compound is active, that is binds to a given protein. We used ten different proteins and corresponding sets of molecules with known (empirically tested) activity. Each compound was represented as the fixed length bit sequence using the SMARTS patterns [33], which is one of the commonly used fingerprints (data representations) in such tasks [8]. These gives us ten different binary datasets, summarized in Table 4.

protein	d	n	$ X_+ $	$ X_- $
5-HT _{2A}	82	2686	1835	851
5-HT ₆	109	1831	1490	341
5-HT ₇	108	1043	704	339
cathepsin	116	1188	245	943
D ₂	137	6215	3342	2873
hERG	130	4928	496	4432
HIV integrase	130	1015	101	914
HIV protease	134	4052	3155	897
M ₁	123	1697	759	938
SERT	129	5231	3559	1672

Table 4: Summary of cheminformatics datasets used in tests. SubFP [8] is used for molecules representation.

Similarly to the previous experiments, we compare MELC with fitted γ parameter ($\gamma \in \{0.1, 0.2, \dots, 1.1\}$) with balanced linear SVM with fitted C . Greedy gradient optimization is performed from the set of starting points consisting of

random points (uniformly selected from the unit sphere), solution of balanced SVM with $C = 1$, and solution of perceptron. The model with highest D_{CS} value is selected. Conducted experiments, summarized in Table 5, show that our method is a competitive model for this kind of data. It is clear that for some proteins (like 5-HT_{2A} or cathepsin, see Figure 2) the internal data geometry can be better exploited using multithreshold linear classifier. Namely such model can detect, contrary to single threshold linear model, some underrepresented classes of active molecules which can be of high importance in the the search for new proteins’ ligands.

protein	MCC		WAC	
	MELC	SVM-B	MELC	SVM-B
5-HT _{2A}	0.434 [2.8]	0.379	0.725 [2.8]	0.703
5-HT ₆	0.604 [3.0]	0.593	0.835 [3.0]	0.834
5-HT ₇	0.464 [9.8]	0.435	0.735 [9.8]	0.723
cathepsin	0.530 [1.0]	0.476	0.796 [1.0]	0.779
D ₂	0.441 [1.0]	0.442	0.720 [1.2]	0.721
hERG	0.320 [1.5]	0.304	0.740 [1.3]	0.738
HIV integrase	0.543 [4.6]	0.515	0.834 [1.1]	0.835
HIV protease	0.501 [1.0]	0.493	0.782 [1.0]	0.782
M ₁	0.536 [3.6]	0.532	0.769 [3.6]	0.766
SERT	0.439 [1.0]	0.438	0.734 [1.4]	0.733

Table 5: Summary of results for cheminformatics data.

By examining scores for some particular folds we can see that despite similarities, MELC and SVM have difficulties in classifying different datasets. In particular one can see on Figure 11 that for 5-HT_{2A} dataset, third fold was the hardest one (in terms of MCC) for SVM while the same data seemed as easy for MELC as the first or the second one. This shows that resulting models are indeed different.

Naturally, better overall results could be obtained using kernelized SVM (with RBF kernel), although it would lead to creation of very complex models (number of support vectors for these datasets varies between 1000 and 2000). As the result, constructed classifier is big and slow (it consists of about 100,000 numbers, and requires thousands of exp evaluations), while at the same time MELC builds very light model, consisting of about $d + 3$ numbers ($d + 10$ in case of 5-HT₇). It is an important factor, as speed of a resulting model is an important aspect for actual applications of compounds activity classifiers, which should be able to process huge databases of possible molecules.

We have also checked how many iterations of *binsearch* is required to build a good (close to the density based) k -threshold linear classifier. For considered datasets, performing just one iteration (placing the threshold in the middle between two points projections) led to very similar results (see Table 6). Performing five iterations led to exactly the same scores as achieved with density

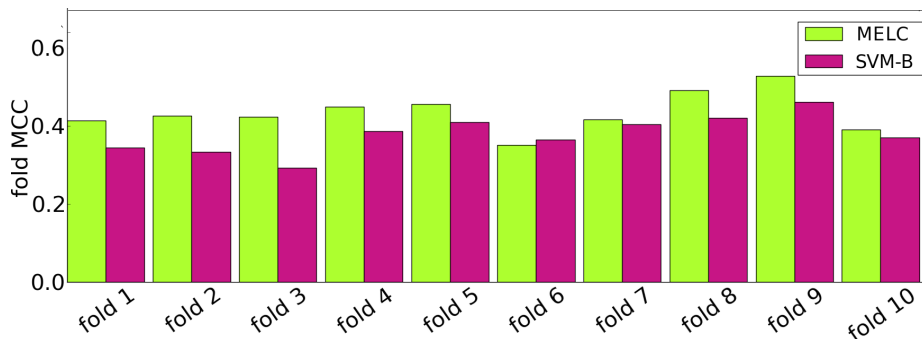


Figure 11: Matthew’s Correlation Coefficient for each fold of 5-HT_{2A} dataset using MELC and balanced SVM.

based model.

protein	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
5-HT _{2A}	0.001	0	0	0	0
5-HT ₆	0.001	0	0	0	0
5-HT ₇	0.004	0.003	0.003	0.001	0
cathepsin	0	0	0	0	0
D ₂	0.001	0.001	0.001	0	0
hERG	0.001	0.001	0.001	0.001	0
HIV integrase	0.003	0.003	0.002	0.002	0
HIV protease	0	0	0	0	0
M ₁	0	0.001	0	0	0
SERT	0	0	0	0	0

Table 6: Differences between MCC scores of density based classifier and k -threshold after i iterations of binsearch. Differences for WAC were even smaller

To sum up, the evaluation on the real, cheminformatics dataset lead to the following conclusions regarding proposed model:

- obtained results are (in most cases) better than those obtained by balanced SVM,
- resulting model has the same complexity as linear models (and rows of magnitude smaller than kernelized ones),
- internal data geometry of chemical compounds can be better exploited using multithreshold model,
- multithreshold structure might lead to detection of underrepresented active/inactive compounds families,

- just a few iterations of binsearch are required to convert a density based method to actual multithreshold function.

7 Conclusions

In this paper we presented a novel multithreshold classification method based on Renyi’s quadratic entropy. Proposed model is based on search for the best linear projection on \mathbb{R} in terms of Cauchy-Schwarz divergence of kernel estimation of the data projection. We showed its theoretical justification and properties, including scale invariance and relations to the largest margin SVM classifier.

We proposed a simple, gradient based constrained optimization method for the construction of density-based classifier. However, it remains an open issue how to efficiently optimize it, as outlined approach has high computational complexity. We also showed how such a classifier can be efficiently converted to the k -threshold linear classifier.

During evaluation we studied how proposed model behaves on UCI binary datasets as well as real data coming from cheminformatics. In most cases, MELC behaved better than balanced SVM in terms of balanced evaluation measures (WAC and MCC). We also investigated existence of correlation between our criterion and the generalization error. Obtained results support our claim that proposed method performs structural risk minimization.

Acknowledgments

This work was partially founded by National Science Centre Poland grant no. 2013/09/N/ST6/03015.

The authors would like to thank Igor Podolak, Phd for his invaluable contribution to our work, discussions, suggestions and criticism. We would also like to thank Sabina Smusz, MSc from Institute of Pharmacology, Polish Institute of Science for providing access to the cheminformatics data and sharing knowledge regarding compounds’ activity prediction. Finally, we would like to thank Daniel Wilczak, Phd for access to the Fermi supercomputer which made the numerous experiments possible.

References

- [1] G.-B. Huang, D. H. Wang, and Y. Lan, “Extreme learning machines: a survey,” *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 107–122, 2011.
- [2] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [3] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

- [4] R. Cao, A. Cuevas, and W. Gonzalez Manteiga, "A comparative study of several smoothing methods in density estimation," *Computational Statistics & Data Analysis*, vol. 17, no. 2, pp. 153–176, 1994.
- [5] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.
- [6] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.
- [7] M. Anthony, "Generalization error bounds for threshold decision lists," *The Journal of Machine Learning Research*, vol. 5, pp. 189–217, 2004.
- [8] S. Smusz, R. Kurczab, and A. J. Bojarski, "The influence of the inactives subset generation on the performance of machine learning methods," *Journal of cheminformatics*, vol. 5, no. 1, pp. 1–8, 2013.
- [9] R. Takiyama, "Multiple threshold perceptron," *Pattern Recognition*, vol. 10, no. 1, pp. 27–30, 1978.
- [10] S. Olafsson and Y. S. Abu-Mostafa, "The capacity of multilevel threshold functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 2, pp. 277–281, 1988.
- [11] M. Anthony, *Learning multivalued multithreshold functions*. Citeseer, 2003.
- [12] K. Huang, H. Yang, I. King, and M. R. Lyu, "Maxi-min margin machine: learning large margin classifiers locally and globally," *Neural Networks, IEEE Transactions on*, vol. 19, no. 2, pp. 260–272, 2008.
- [13] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine learning*, vol. 37, no. 3, pp. 277–296, 1999.
- [14] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.
- [15] J. C. Principe, *Information theoretic learning*. Springer, 2000.
- [16] J. C. Principe, R. Jenssen, and S. Rao, "Clustering with itl principles," in *Information theoretic learning*. Springer, 2000, pp. 263–298.
- [17] J. C. Principe, S. Rao, D. Erdogmus, D. Xu, and K. I. Hild, "Self-organizing itl principles for unsupervised learning," in *Information theoretic learning*. Springer, 2000, pp. 263–298.
- [18] J. M. Santos, L. A. Alexandre, and J. M. de Sá, "The error entropy minimization algorithm for neural network classification," in *International Conference on Recent Advances in Soft Computing*. Citeseer, 2004, pp. 92–97.
- [19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

- [20] N. Timm, *Applied multivariate Analysis*. Springer Text in Statistics, 2002.
- [21] P. Drineas and M. W. Mahoney, “On the nyström method for approximating a gram matrix for improved kernel-based learning,” *The Journal of Machine Learning Research*, vol. 6, pp. 2153–2175, 2005.
- [22] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: theory and applications,” *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.
- [23] C. Hegde, M. B. Wakin, and R. G. Baraniuk, “Random projections for manifold learning.” in *NIPS*, vol. 7, 2007, p. 59.
- [24] S. S. Haykin, *Neural networks and learning machines*. Pearson Education Upper Saddle River, 2009, vol. 3.
- [25] B. Karlsson, *Beyond the C++ standard library: an introduction to boost*. Pearson Education, 2005.
- [26] A. Asuncion and D. Newman, “Uci machine learning repository,” 2007.
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [28] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [29] F. Hammann, H. Gutmann, U. Baumann, C. Helma, and J. Drewe, “Classification of Cytochrome P 450 Activities Using Machine Learning Methods,” *Molecular Pharmaceutics*, vol. 33, no. 1, pp. 796–801, 2009.
- [30] X. Zhang, X. Liu, and Z. J. Wang, “Evaluation of a set of new orf kernel functions of svm for speech recognition,” *Engineering Applications of Artificial Intelligence*, vol. 26, no. 10, pp. 2574–2580, 2013.
- [31] A. Subasi, “Classification of emg signals using pso optimized svm for diagnosis of neuromuscular disorders,” *Computers in biology and medicine*, vol. 43, no. 5, pp. 576–586, 2013.
- [32] P. Van Kerm, “Adaptive kernel density estimation,” *Stata Journal*, vol. 3, no. 2, pp. 148–156, 2003.
- [33] C. W. Yap, “Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints,” *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.